W492–W499 *Nucleic Acids Research, 2004, Vol. 32, Web Server issue*
DOI: 10.1093/nar/gkh443

GOAL: automated Gene Ontology analysis of expression profiles

Stefano Volinia[1,2,*], Rita Evangelisti[1], Francesca Francioso[1], Diego Arcelli[3], Massimo Carella[2,4] and Paolo Gasparini[2,5]

[1]Laboratory of Functional Genomics and Telethon Facility—Data Mining for Analysis of DNA Microarrays, Department of Morphology and Embryology, Università degli Studi, Via Fossato di Mortara 64/b, 44100 Ferrara, Italy, [2]TIGEM, Telethon Institute of Genetics and Medicine, Naples, Italy, [3]Istituto Dermopatico dell'Immacolata, Rome, Italy, [4]IRCCS CSS Servizio di Genetica Medica, S. Giovanni Rotondo, Foggia, Italy and [5]Dipartimento Di Patologia Generale, Seconda Università, Naples, Italy

Received December 23, 2004; Revised and Accepted April 20, 2004

ABSTRACT

One of the most common problems encountered while deciphering results from expression profiling experiments is in relating differential expression of genes to molecular functions and cellular processes. A second important problem is that of comparing experiments performed by different labs using different microarray platforms, or even unrelated techniques. Gene Ontology (GO) is now used to describe biological features, since GO terms are associated with genes, to overcome the apparent distance between expression profiles and biological comprehension. Here we describe the development, implementation and use of GOAL (Gene Ontology Automated Lexicon), a web-based application for the identification of functions and processes regulated in microarray and SAGE (serial analysis of gene expression) experiments. We applied GOAL to a range of experimental datasets related to different biological problems, including cancer and the cell cycle. By using GOAL, reported and novel relevant processes were identified in a number of experiments by our collaborators and by us. Different datasets could also be compared with each other to define conserved functional modules. GOAL allows a seamless and high-level analysis of expression profiles and is implemented as a free WWW resource (http://microarrays.unife.it).

INTRODUCTION

Biomedical research over the last decades has made tremendous progress in the understanding of biology and medicine. The sequencing of the genomes of human, mouse and other organisms, in combination with high-throughput procedures such as those based on microarray and SAGE (serial analysis of gene expression) techniques, has meanwhile started yielding massive amounts of data, often stored in public databases. However, full utilization of these data and their integration with existing knowledge from different domains has to be facilitated by automation towards a systematic representation of knowledge. Recently, the Gene Ontology (GO) Consortium (http://www.geneontology.org) has developed a systematic and standardized nomenclature for annotating genes in various organisms, including human (1,2). Using the three main ontologies molecular function, biological process and cellular component, a significant number of genes in yeast, *Drosophila*, mouse and human have been annotated (3). GO assignments have also recently been applied to a dataset representing the complete human proteome, using a combination of electronic mappings and manual curation, by the GO annotation project (GOA) at the European Bioinformatics Institute (EBI), including SwissPROT and TrEMBL (4). Interpretation of results from high-throughput-generated expression profiles is hampered by the very large amount of data obtained in a typical experiment, and literature-based algorithms have been devised to gain functional insight (5). The GO project provides the information necessary for the interpretation of the expression patterns once each gene is associated with its related GO term(s), describing function, processes and cellular component. The availability of GO annotations for a significant number of genes from different organisms presents an opportunity to examine the cellular localization, molecular function and involvement in a biological process of each of them through the multiple and hierarchical structure of GO.

For this reason we have developed a resource, GOAL (Gene Ontology Automated Lexicon), for automated and streamlined functional analysis of expression profiles. We use it here to

*To whom correspondence should be addressed. Tel: +39 0532 291714; Fax: +39 0532 291533; Email: s.volinia@unife.it demonstrate the efficient and automated assignment of GO terms to whole expression profiles generated by a number of labs studying different biological problems and to detect those GO terms which are significantly regulated.

This WWW resource automatically generates and evaluates scoring of molecular functions, biological processes and cellular components from the results of an expression profiling experiment. The resource's tools can be applied to different biological problems, from multiple sample groups to time course experiments. Permutation analysis is performed to define *P*-values and false detection rates (FDRs) within each dataset. GOAL can be used for analysis of cDNA microarrays, oligonucleotide microarrays and SAGE.

MATERIALS AND METHODS

Data collection

We used SOURCE (6) to download the GO terms associated with each clone measured by the microarray and the SAGE experiments described in this paper via the clone ID or the Genbank accession number. A MySQL database was used to store the relative tables. For other platforms, such as Affymetrix, we used the Genbank accession number related to each Probe Set. Affymetrix Probe Set Ids are linked to the respective Genbank accession number in a separate table. Since the purpose of GOAL is to act as a public resource for the scientific community, our policy is to store in the GOAL MySQL database all the annotations present in the (Gene Expression Omnibus) GEO (7) and Array-Express public databases (8) related to probes or reporters, i.e. oligonucleotides, expressed sequence tags (ESTs). The GOAL database contains the GO terms representing the human and mouse Affymetrix chips and the Rosetta Inpharmatics 25K chip, NEN, Unigem by Incyte, Agilent, CODElink (Amersham), MWG and UCSD Biogem (human and mouse). The NHGRI and OCI layouts (both human and mouse cDNA) were entered into the GOAL database as part of the 40K Research Genetics library (human) and the mouse 15K NIA collection. The Stanford human 40K set is also annotated in the GOAL database, after downloading gene lists from the Stanford Microarrays Database (9). Annotations for other human microarray platforms and for some SAGE datasets stored in the GEO database at the NCBI were also entered into the GOAL database. A total of >50 000 different reporters are currently annotated in the database and allow translation of Genbank accession numbers, IMAGE clone IDs and Affymetrix Probe Sets into GO terms. Complete coverage of the GEO and Array-Express databases is being pursued.

Algorithms

The Goal resource is essentially a combination of two tasks. The first task consists of data input and association of the expression table to statistics, and then to Unigene clusters and GO terms. The second task is the actual GO analysis script, which calculates the scores for each Unigene cluster, and then for each GO term. The statistical scores are calculated for the different ESTs of the same Unigene cluster, and for each different Unigene cluster associated with the same GO term. Finally, *P*-values and false detection rates for properly describing the statistical significance of the results are calculated.

There are two possible GOAL analysis routes: one starting from an expression table, and the other from expression statistics, i.e. a table containing probe IDs and related statistical scores.

Analysis of expression statistics

A GOAL analysis route can therefore start with the submission of a table containing the reporter IDs and the related scores from a user-performed statistical analysis. Acceptable data include logged or unlogged intensity ratios from a two-channel experiment or a one-channel table including at least a reference baseline sample or samples group. Units accepted for SAGE datasets include TPM (tags per million), the standard measure for this technique. Examples of the different file formats are hyperlinked to the dataset submission forms. Gene lists can be identified by either Genbank accession numbers, Affymetrix Probe Set IDs or IMAGE clone IDs. SAGE gene identifiers are Genbank accession numbers. Preprocessing steps, such as background subtraction, quality control and normalization, need to be applied to the data prior to submission. Since only different genes must be scored, the script reduces the repetitions due to different clones/ESTs pointing to the same Unigene cluster by transparently calculating the mean, median or trimmed mean score of each Unigene cluster; that is, for each clone ID or Genbank entry the corresponding Unigene cluster is retrieved. The parsing script associates GO terms with each EST in the uploaded file, which includes IMAGE clone IDs or Genbank accession numbers, and the relative scores. After the GO terms have been identified for each EST featured in the expression table, GOAL calculates the average score, and if possible the *P*-value for each of the GO terms in the experiment under investigation. False detection rates are also calculated to measure multiple testing effects (10). A web page is then returned to the user with all the results and statistics for the dataset, including the shapes of the score distributions generated by the permutation analysis for different *k* (*k* being the number of genes associated with a GO term). The output from GOAL also contains the score and *P*-value for each significant annotated Unigene cluster. *P*-values can be chosen at three different stringency levels ($P < 0.05$, $P < 0.01$ and $P < 0.001$). The score distribution is generated for each submitted dataset and for each *k*-tuple. Permutation tests are automatically performed on each submitted dataset by randomly permuting the score elements. Only the scores of genes and ESTs with GO annotation are used. A graphical output page containing false detection rates and the observed/expected distributions is provided to the user. If a reporter ID is not present in the GOAL database, the script reveals its absence to the user and SOURCE is automatically queried. If a positive response is obtained, the GO annotation is automatically entered into the GOAL database. Nevertheless, regular GOAL database updates ensure that genes/ESTs and their relative GO terms are coordinated with external databases and relieve the user from any updating task.

Data submission

The first task, data submission via file uploading from a WWW interface, can be performed by two different applications, representing alternative procedures: one which calculates the statistics from an inputted expression table, and one which accepts gene names and scores from a user's precomputed statistics, as detailed above. The first procedure ensures a swift GO analysis starting directly from the expression table and is limited for the moment to a *t*-test comparison, while the second, 'open' procedure allows the use of any external statistical method that the user might consider appropriate. Nevertheless, in the next GOAL release we expect to add more internal statistical tests besides the *t*-test in order to further facilitate queries for the user. In this paper, both procedures are described using real datasets, and the open procedure using different statistical procedures from the widely used SAM Excel add-in to the Fourier transform.

Submission procedure I: expression table

Starting from an expression table, a Perl script (tScan) computes a statistics, here *t*-test, over two groups (i.e. wild-type and mutant groups) and writes to a temporary text file the Unigene cluster and the GO terms retrieved from the GOAL SQL database. A script variation allows the evaluation on a single group of samples if in a Cy5–Cy3 experiment the wild type is also the common reference sample. In this case, the second group is automatically generated using a user-provided parameter (i.e. the standard deviation measured in an actual experiment by co-hybridizing the same RNA labeled with two cyanines). The second control group is built by generating random values within boundaries of ±3× the standard deviation measured in control experiments when an identical RNA sample is labeled with two different dyes (currently 0.51 in our lab for indirect labeling experiments and CMT–GAP Corning slides). The *t*-score for each row is computed as the average of the *t*-scores obtained in a number of cycles, by default 50. Values in the expression table can be intensity ratios, for a typical two-dyes experiment, or absolute intensities, for a one-channel experiment, and can be entered either as logged or unlogged. An additional table containing the *t*-scores related to each row in the expression matrix, computed for each one of 10 balanced permutations, is written in order to perform permutation analysis within the subsequent DirectGO analysis script. Either all table entries are used (transcriptome-wide analysis) or a *t*-score threshold can be applied to filter out ESTs or genes with low *t*-scores (restricted analysis). Scores are calculated and reported in the tScan output files only for GO-annotated entries. Two different input forms are available, a simple interface and an advanced interface where most parameters are customizable by the user.

Submission procedure II: statistics table

Rather than a whole expression table, as in procedure I, here a gene list associated with a statistics score for each gene is submitted to GOAL via the ScoreScan script. We define the inputted 'gene list plus scores' table as the statistics table. The advantage is total flexibility to the user's need, since any statistical procedure can in principle be used. The user executes the appropriate statistical analysis prior to data submission by using an external application, e.g. SAM, Fourier transformation or Bioconductor. The statistics table is then submitted to GOAL via the ScoreScan web interface. Robust experiment-wise estimate for *P*-values and FDR calculation can be attained by submitting extra scores calculated by permuted groups during the external statistical analysis. A number of these extra score columns can in fact be added to the submitted statistics table, and their number entered in the submission form.

Gene Ontology *P*-values and FDR: DirectGO

In the analysis step following either submission procedure I or II, the DirectGO script reads the output files generated above. In the case of procedure I, tScan, two output files are necessary in order to calculate the scores linked to each GO term and to each Unigene cluster. First, the *t*-scores for all different clone IDs or Genbank entries referenced by the same Unigene cluster are averaged. Finally, scores for each GO term are obtained as the mean, median or trimmed mean of the scores for the different Unigene clusters linked to that GO term. Meanwhile, *P*-values are also attached to each Unigene cluster by comparing the real scores to the *t*-scores distribution obtained from the balanced permutation table. This procedure allows the user to identify the annotated Unigene clusters which are differentially expressed. Although an intermediate step towards GO analysis, this is already a valuable result for the user. The presence of an up-to-date Unigene cluster database allows the user a transparent approach to Unigene cluster statistical analysis, and it is a useful feature of the GOAL resource. Annotated Unigene clusters which are differentially expressed can in fact be promptly identified.

In order to obtain *P*-values and FDR a distribution is obtained by performing a permutation analysis. Robust experiment-wise estimates are obtained by using data produced by tScan during balanced permutation of the expression table. This procedure should guarantee that *P*-values and FDR are not affected by large gene expression differences in the sample groups, as might happen when comparing cancer with normal tissues, where as many as 20% of the genes could be differentially expressed. *P*-values and FDR are calculated from permutations specific to each different k, k being the number of different Unigene clusters pointing to GO terms; e.g., calcium-sensitive guanylate cyclase activator is a $k = 2$ GO term, being associated in a dataset with two different Unigene clusters, while cyclin-dependent protein kinase might be in the same dataset a $k = 4$ GO term, being linked to four Unigene clusters. k ranges from a minimum of 2 to a maximum of 9. Any value above 9 is included in the ninth class (Figure 1).

In the case of submission procedure II, the distribution shape needs to be specified by the user, i.e. two- or one-tailed, right or left significance when one-tailed (Figure 2). Another user-defined parameter is the number of columns, if any, related to the scores from balanced permutations (a maximum of 10).

RESULTS AND DISCUSSION

GOAL enables a statistical approach to GO analysis. GOAL allows evaluation of *P*-values and FDR, both important parameters for establishing the practical usefulness of selected genes and function. There are two alternative data-entry procedures, one which computes statistics on an expression table and the other which uses the output from an external statistical

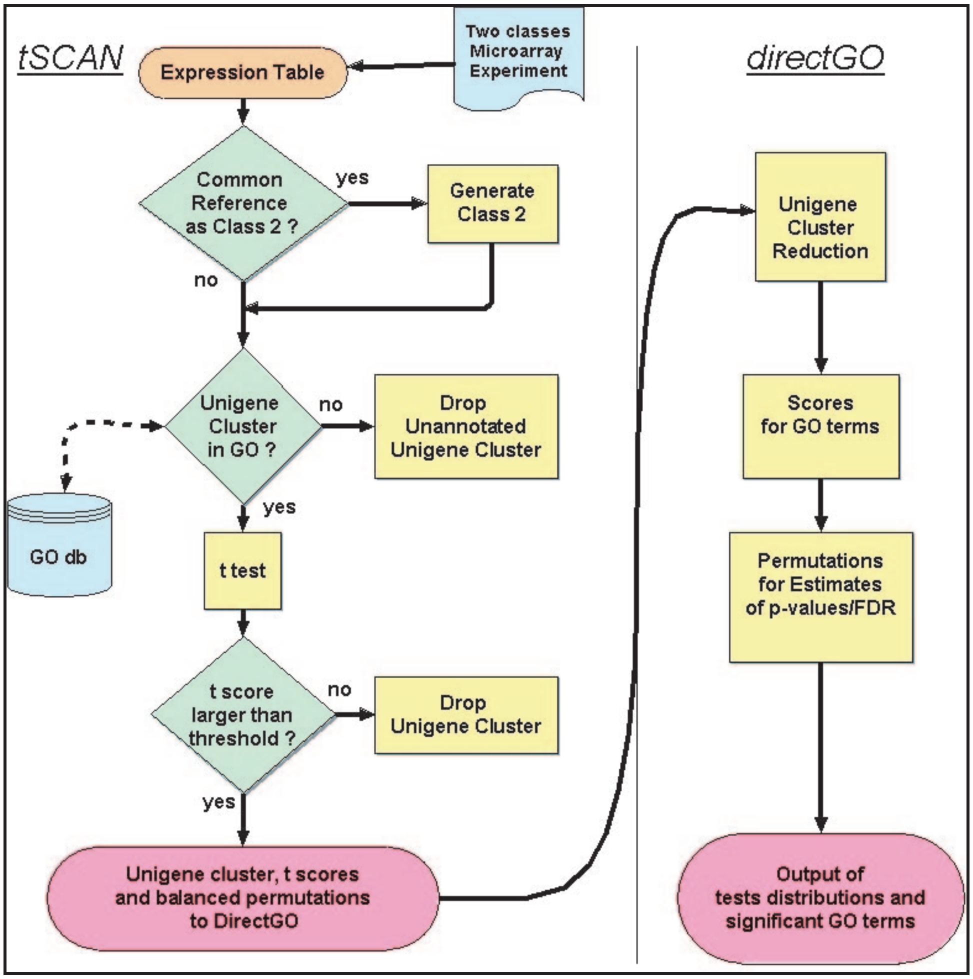

tSCAN
Expression Table
Two classes Microarray Experiment
Common Reference as Class 2 ?
yes
Generate Class 2
no
Unigene Cluster in GO ?
no
Drop Unannotated Unigene Cluster
GO db
yes
t test
t score larger than threshold ?
no
Drop Unigene Cluster
yes
Unigene cluster, t scores and balanced permutations to DirectGO
directGO
Unigene Cluster Reduction
Scores for GO terms
Permutations for Estimates of p-values/FDR
Output of tests distributions and significant GO terms

Figure 1. Flow chart for a two-class analysis. When a two-class analysis (i.e. treated versus untreated samples) is performed, the tSCAN script calculates *t*-scores for each annotated EST or reporter in the expression table. If the second class, i.e. the untreated control, is used in a two-dyes experiment as the common reference, the script generates automatically a second class of identical size with the relative control values (intensity log ratios). *t*-scores from tSCAN and the relative results from balanced permutations are then inputted into DirectGo. DirectGO assigns replicated or redundant gene fragments to each Unigene cluster and calculates a group score by using mean, trimmed mean or median. Group scores are similarly computed for each GO term represented by the Unigene clusters. The balanced permutations method is applied to simulate the null distribution in order to obtain false detection rates and confidence limits.

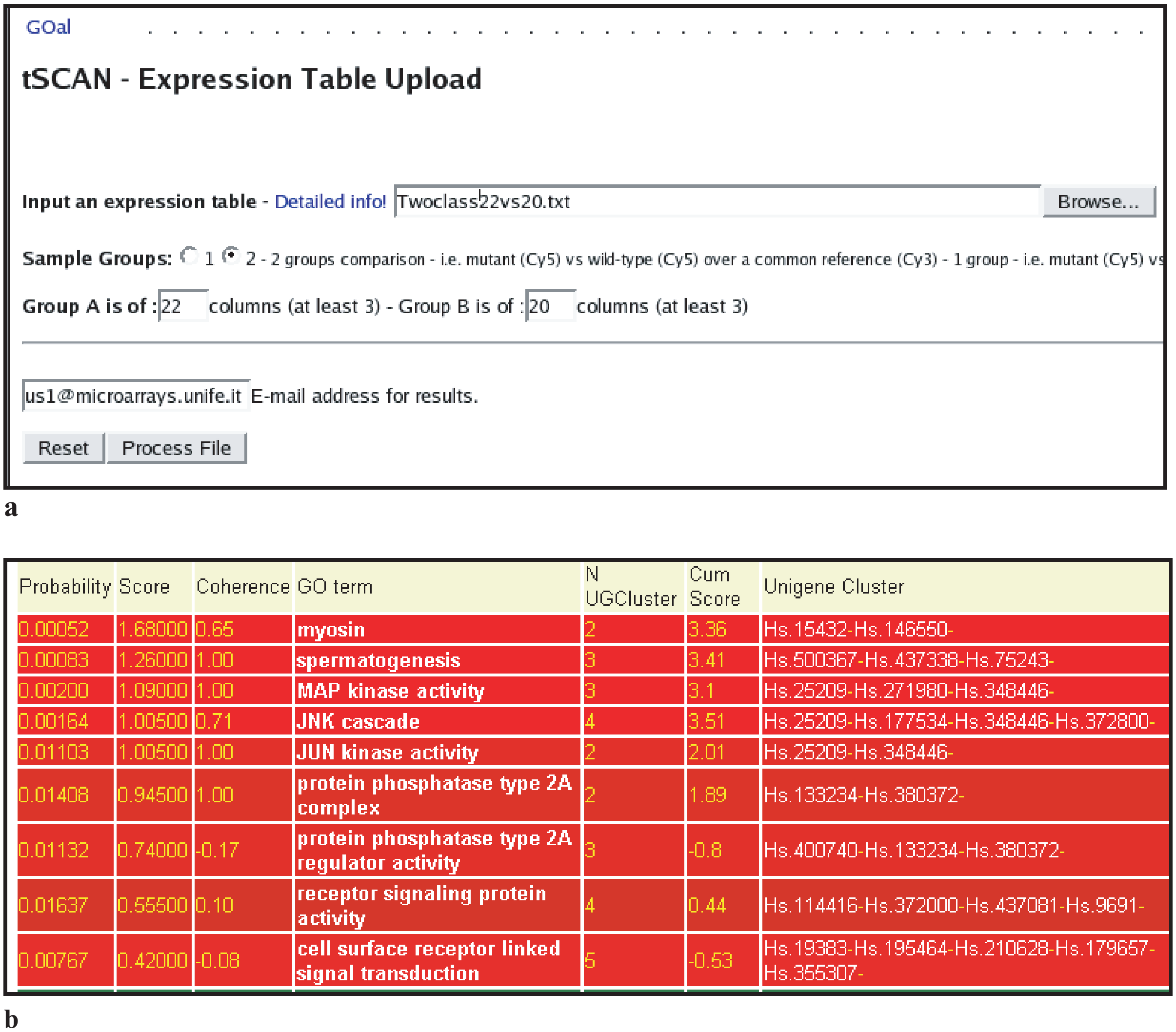
package, such as Bioconductor (11), SAM (12) or GeneSpring (http://www.genespring.com).

In the first instance, for example, a two-class comparison can be performed by using a *t*-test approach coupled to permutation analysis on an expression table. As a special case, a one-class analysis can be performed when the common reference used in a two-dyes experiment is itself one of the two classes, i.e. disease sample versus control. In order to produce efficient experiment-wise estimates of significance levels in the subsequent GO terms procedure, a set of 10 score columns is generated by balanced permutations and saved for further analysis.

In the alternative procedure, any external statistical application can be used; the GOAL input is thus a table containing the score associated with each gene on the chip, e.g. a Fourier transform. The submitted table contains at least two columns, i.e. the gene ID and the statistical test score columns. Extra columns with the results from balanced permutations, in order to allow GOAL efficient FDR evaluation, can be added into the table.

After data upload, a web page is generated with a list of all GO terms differentially regulated in the experiment, according to the scoring procedure applied. To provide a *P*-value for each GO term and FDR, a permutation analysis (13) is performed on the submitted dataset by permuting each row within the score column, or, if present, within each extra score column calculated from balanced permutations. In order to assess the number of permutations needed for robust generation of *P*-values and determination of FDR, a number of datasets were used and a range of permutations were performed within each dataset. Varying the number of cycles, and depending on the dimension of the expression tables, we generated permutations with up to 2×10^6 total data. It appears that $\sim 5 \times 10^5$ total scores are sufficient for optimal FDR evaluation, since by increasing the number of cycles and subsequently the amount of data generated, the shape of the test distribution does not seem to be affected (Figure 3).

An important side-effect of using GOAL is the automated conversion of ESTs/oligonucleotides to Unigene clusters. The vast majority of packages for expression profile analysis



GOal

tSCAN - Expression Table Upload

Input an expression table - Detailed info! Twoclass22vs20.txt Browse...

Sample Groups: ○ 1 ● 2 - 2 groups comparison - i.e. mutant (Cy5) vs wild-type (Cy5) over a common reference (Cy3) - 1 group - i.e. mutant (Cy5) vs

Group A is of : 22 columns (at least 3) - Group B is of : 20 columns (at least 3)

us1@microarrays.unife.it E-mail address for results.

Reset | Process File

Probability	Score	Coherence	GO term	N UGCluster	Cum Score	Unigene Cluster
0.00052	1.68000	0.65	**myosin**	2	3.36	Hs.15432-Hs.146550-
0.00083	1.26000	1.00	**spermatogenesis**	3	3.41	Hs.500367-Hs.437338-Hs.75243-
0.00200	1.09000	1.00	**MAP kinase activity**	3	3.1	Hs.25209-Hs.271980-Hs.348446-
0.00164	1.00500	0.71	**JNK cascade**	4	3.51	Hs.25209-Hs.177534-Hs.348446-Hs.372800-
0.01103	1.00500	1.00	**JUN kinase activity**	2	2.01	Hs.25209-Hs.348446-
0.01408	0.94500	1.00	**protein phosphatase type 2A complex**	2	1.89	Hs.133234-Hs.380372-
0.01132	0.74000	-0.17	**protein phosphatase type 2A regulator activity**	3	-0.8	Hs.400740-Hs.133234-Hs.380372-
0.01637	0.55500	0.10	**receptor signaling protein activity**	4	0.44	Hs.114416-Hs.372000-Hs.437081-Hs.9691-
0.00767	0.42000	-0.08	**cell surface receptor linked signal transduction**	5	-0.53	Hs.19383-Hs.195464-Hs.210628-Hs.179657-Hs.355307-

Figure 2. Examples from the WWW interface for a two-class analysis. (**a**) An expression table containing a total of 42 samples and 2001 genes is uploaded. The size of each group is entered into the form. The first group starts at the first sample column. (**b**) The output table with the significant upregulated GO terms (total false detection rate 0.19). The median was used to evaluate GO term scores, and the *P*-value threshold was 0.05; 50 permutations were performed to generate a null distribution with over a million GO term scores.

in fact use a single probe/target approach, i.e. selects the differentially expressed cDNA clones or oligonucleotide. GOAL, however, uses the latest Unigene build in order to compute the mean score for all the ESTs/oligonucleotides related to that Unigene cluster. This procedure, necessary to associate GO terms with genes, leads to the reduction of the complexity of the dataset.

Examples of GOAL application to selected published datasets—namely, healthy blood variation (14), diffuse large B-cell lymphoma transformation (15), renal cancer (16), soft tumors (17), lung adenocarcinoma (18) and breast cancer (19)—are outlined in the Web supplement (http://microarrays.unife.it/GOAL/). SAGE (20) datasets were used by entering into the expression table the TPM values as retrieved from GEO.

Transcriptome wide and restricted Gene Ontology analysis

Besides the two data-entry procedures, alternative routes to GO analysis can be followed by the user. In one instance, only those genes which are differentially expressed within the experiment can be used to infer GO results. This method, which we call 'restricted' because it takes into consideration only the subset of genes which are regulated, allows faster analysis and the evaluation of those functions solely related to the pool of regulated genes. This algorithm is similar to that used by most GO applications but, by being restricted to the subset of differentially expressed genes, might miss a fraction of the cell-wide regulated functions and processes. For example, an upregulated process might result from the coordinated upregulation of a number of genes, even though all of them have scores slightly below the significance threshold.

A second path that can be followed by the user is when all the genes in an expression profile are considered for GO analysis. In this way, information is gathered from all the mRNAs measured in the experiment, not only from differentially regulated ones. This method, which we call 'transcriptome-wide', is slower and might yield somewhat different results when compared with the 'restricted' approach. For example,

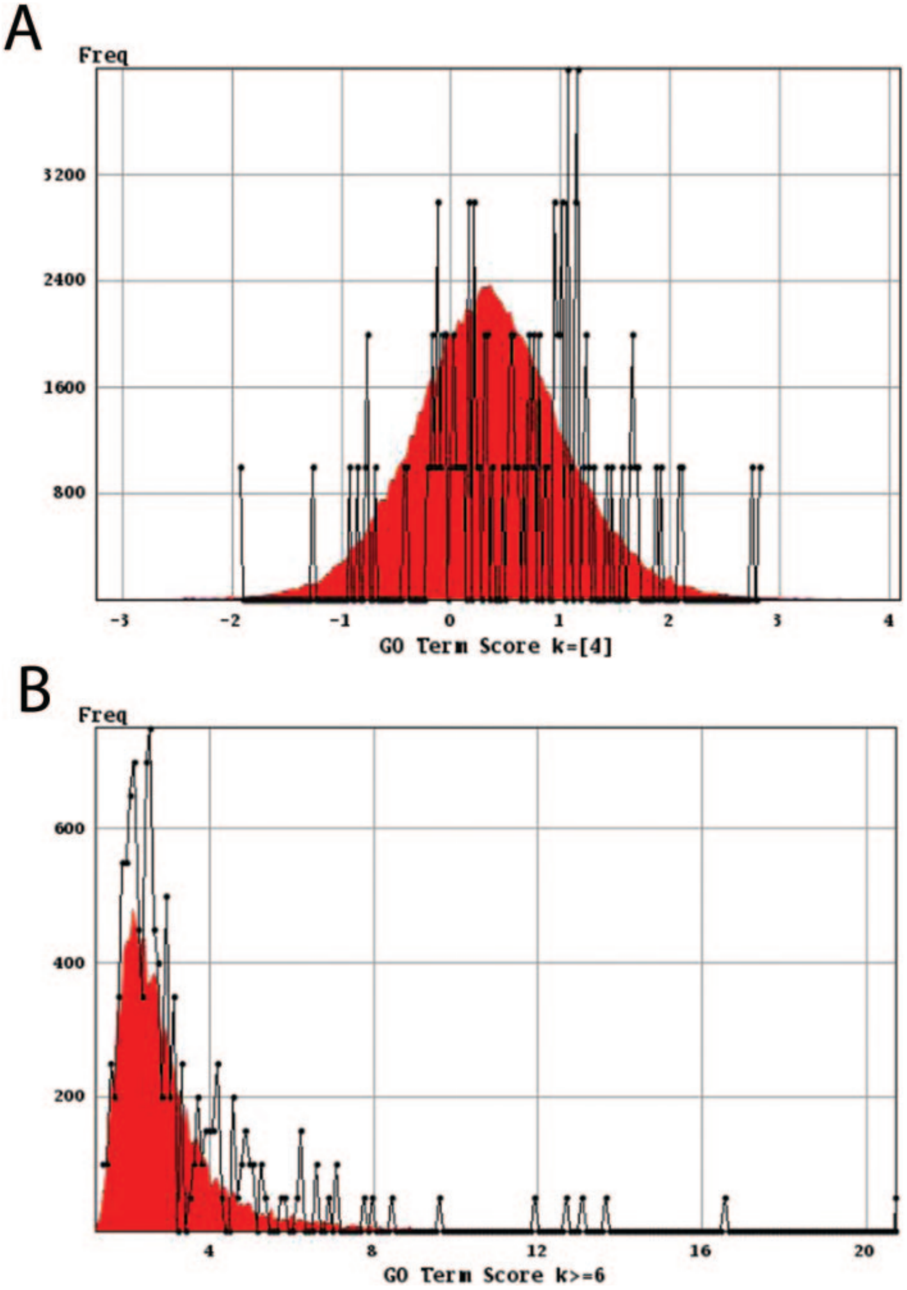


Figure 3. Evaluation of Gene Ontology terms significance. (**A**) *t*-score distribution and the relative test distribution generated by permutation analysis for $K = 4$. In order to compute the *P*-values, 100 permutations were performed by using each of 10 *t*-scores columns obtained by balanced permutations of the actual expression table. A total of $>1.7 \times 10^6$ *t*-scores were generated. For $k = 4$ in this experiment, 102 GO terms were scored in the actual expression table, with a mean of 0.6 and a standard deviation of 0.81. In the permutation analysis, data for a total of 101 909 GO terms were generated, with a mean of 0.32 and a standard deviation of 0.69. The two-tailed confidence limits with $P < 0.05$ are at -1.00 for the downregulated GO terms and 1.73 for the upregulated GO terms. The black peaks correspond to the GO terms in the experiments, while in red is displayed the curve obtained from the permutation analysis. The scale on the left is for the permutation analysis. (**B**) Fourier-score distribution and the relative test distribution generated by permutation analysis for $k \geqslant 6$. These plots are from the cell cycle experiment by Whitfield *et al.* [(31), Web Supplement], using Fourier transforms in order to detect genes with a periodic expression pattern during the cell cycle. In the expression statistics 30 000 ESTs are present and 6910 are GO annotated. In order to compute the *P*-values, 50 permutations were performed using the entered data directly. A total of $>6.9 \times 10^4$ data were generated. For $k \geqslant 6$ in this experiment, 22 GO terms out of 223 scored in the dataset were significant ($P < 0.05$) with an FDR of 0.5. The one-tailed positive distribution confidence limit derived from the permutation analysis for the cell-cycle-correlated GO terms ($P < 0.05$) is at 5.22 (mean of 2.84 and a standard deviation of 1.24). Note that the single gene significance threshold applied by the authors was 3.5. The black peaks correspond to the GO terms in the experiments, while in red is displayed the curve obtained from the permutation analysis. The scale on the left is for the permutation analysis.

invariant genes might affect the differential regulation of a cellular process, even when other genes related to the same process are differentially transcribed. Or, as explained above, all the genes representing a molecular function might be upregulated, but just below the significance threshold used for gene selection; nevertheless, when a transcriptome-wide analysis is performed the corresponding GO term will be selected as significantly upregulated.

CONCLUSION

To speed up and facilitate the comprehension of gene expression changes measured by microarray and other high-throughput techniques an automated analysis of experimental results is necessary. Comparison of different experimental datasets is also of prime importance but as yet difficult to attain. If microarray layouts are different, and this is currently true even for the different releases of commercial providers, or different arraying procedures are used (e.g. cDNA spotting, oligonucleotide photolithography), the task of comparing datasets can be very laborious. Moreover, in complex genomes, such as the human genome, different isoforms with identical enzymatic activity or molecular function are present. To compare different experimental datasets in a typical high-throughput fashion, to correlate genes to functions, an appropriate route might be that of using the GO annotation. GO analysis of microarray results, typically a list of gene IDs, can be slow and tedious, when performed using a manual or semi-automated approach. For example, the newly developing Bioconductor suite (11) contains some specialized packages concerned with GO annotation of genes, but automated GO analysis still does not fully appear in the packages.

A number of dedicated GO applications have recently been developed. Amongst them are Onto-Express (21–23), GenMapp (24), EBI's Expression Profiler GO browser (25), GoMiner (26), ChipInfo (27), NetAffx (28) and FatiGO (29), which are capable of associating at least portions of expression profiles to GO terms.

Before GOAL, and with the exception of GenMAPP and MAPPFinder (30), none of the above mentioned applications had been devised for a holistic approach to functional analysis of expression profiles. Moreover, statistical evaluation of the GO analysis needed to be improved; in particular FDR needed to be calculated. Therefore, we developed GOAL, a web-based application, to perform holistic GO analysis and identify regulated GO terms in microarray and SAGE series. From a typical experiment as many as several thousand different GO terms can be scored by GOAL. We designed GOAL in order to take full advantage of the very large amount of information present in the expression profiles. In fact, to comprehend expression profiles at a transcriptome-wide level, those genes whose expression changes abruptly are as important as the invariant genes. Additionally, genes belonging to the same pathway might undergo opposite changes in expression. But for the investigator who wishes to concentrate only on the significantly affected genes, GOAL can perform a restricted analysis, where only those genes scoring above a predefined threshold are taken forward to GO analysis. The results of these two approaches might be very similar or might differ to a certain extent, and the user has the option to follow the most suitable approach. Restricted analysis is faster than transcriptome-wide analysis and it might reveal the best choice for a first approach to the analysis. Results are visualized in a user-friendly fashion, i.e. red when overexpressed, green when underexpressed. Two-tailed or one-tailed distributions can be evaluated.

As an important side-effect of the GO analysis, GOAL allows the identification of single genes whose expression is significantly altered in an experiment. Unlike most analysis programs, average scores are first calculated for identical ESTs spotted on the chip, and then for ESTs belonging to the same Unigene cluster. GO annotations are displayed in the final output of the most significant genes, in addition to hyperlinks to relevant external databases. When possible the *P*-values, computed by balanced permutations, for each significant gene are present in the results—information not found in other applications solely devoted to identification of differential gene expression. Unlike most other available packages for detection of differentially expressed genes, GOAL works by considering Unigene clusters, rather than gene fragments, such as the ESTs or oligonucleotides arrayed onto a chip. Long-term GOAL upgrading and database maintenance will be performed by the staff of the Functional Genomics Laboratory and of the 'Data Mining for Analysis of DNA Microarrays' Telethon Facility. Since gene identification in the human genome and in other genomes is still a dynamic process, this GOAL feature, in parallel with constant updating of the current Unigene build, is a considerable benefit for the study of annotated genes.

SUPPLEMENTARY INFORMATION

The application and supplementary data can be found at http://microarrays.unife.it.

ACKNOWLEDGEMENTS

The authors thank the Gene Ontology Consortium, Stanford Microarray Database, SOURCE and all the authors of the published datasets used in this study for making their data available. We also thank Dr Sandro Banfi and Dr Hillary Siddons for reading the manuscript. MIUR PRIN, University of Ferrara grants to S.V. and the financial support of Telethon—Italy (Grant no. GTF03012) are gratefully acknowledged.

REFERENCES

1. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T., Harris,M.A., Hill,D.P., Issel-Tarver,L., Kasarskis,A., Lewis,S., Matese,J.C., Richardson,J.E., Ringwald,M., Rubin,G.M. and Sherlock,G. (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nature Genet.*, **25**, 25–29.
2. Harris,M.A., Clark,J., Ireland,A., Lomax,J., Ashburner,M., Foulger,R., Eilbeck,K., Lewis,S., Marshall,B., Mungall,C. *et al.* (2004) Gene Ontology Consortium. The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res.*, **32**, D258–D261.
3. The Gene Ontology Consortium. (2001) Creating the gene ontology resource: design and implementation. *Genome Res.*, **11**, 1425–1433.
4. Camon,E., Magrane,M., Barrell,D., Binns,D., Fleischmann,W., Kersey,P., Mulder,N., Oinn,T., Maslen,J., Cox,A. *et al.* (2003) The Gene Ontology Annotation (GOA) Project: Implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Res.*, **13**, 662–672.
5. Jenssen,T.K., Laegreid,A., Komorowski,J. and Hovig,E. (2001) A literature network of human genes for high-throughput analysis of gene expression. *Nature Genet.*, **28**, 21–28.
6. Diehn,M., Sherlock,G., Binkley,G., Jin,H., Matese,J.C., Hernandez-Boussard,T., Rees,C.A., Cherry,J.M., Botstein,D. and Brown,P.O. *et al.* (2003) SOURCE: a unified genomic resource of functional annotations, ontologies, and gene expression data. *Nucleic Acids Res.*, **31**, 219–223.
7. Edgar,R., Domrachev,M., Lash A.E. (2002) Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.*, **30**, 207–210.
8. Brazma,A., Parkinson,H., Sarkans,U., Shojatalab,M., Vilo,J., Abeygunawardena,N., Holloway,E., Kapushesky,M., Kemmeren,P.,

Lara,G.G. *et al*. (2003) ArrayExpress—a public repository for microarray gene expression data at the EBI. *Nucleic Acids Res*., **31**, 68–71.

9. Gollub,J., Ball,C.A., Binkley,G., Demeter,J., Finkelstein,D.B., Hebert,J.M., Hernandez-Boussard,T., Jin,H., Kaloper,M., Matese,J.C. *et al*. (2003) The Stanford Microarray Database: data access and quality assessment tools. *Nucleic Acids Res*., **31**, 94–96.
10. Francioso,F., Carinci,F., Tosi,L., Scapoli,L., Pezzetti,F., Passerella,E., Evangelisti,R., Pastore,A., Pelucchi,S., Piattelli,A. *et al*. (2002) Identification of differentially expressed genes in human salivary gland tumors by DNA microarrays. *Mol. Cancer Ther*., **1**, 533–538.
11. Dudoit,S., Gentleman,R.C. and Quackenbush,J. (2003) Open source software for the analysis of microarray data. *Biotechniques*, (Suppl.) 45–51.
12. Tusher,V.G., Tibshirani,R. and Chu,G. (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl. Acad. Sci. USA*, **98**, 5116–5121.
13. Carinci,F., Bodo,M., Tosi,L., Francioso,F., Evangelisti,R., Pezzetti,F., Scapoli,L., Martinelli,M., Baroni,T., Stabellini,G. *et al*. (2002) Expression profiles of craniosynostosis-derived fibroblasts. *Mol. Med*., **8**, 638–644.
14. Whitney,A.R., Diehn,M., Popper,S.J., Alizadeh,A.A., Boldrick,J.C., Relman,D.A. and Brown PO. (2003) Individuality and variation in gene expression patterns in human blood. *Proc. Natl Acad. Sci. USA*, **100**, 1896–1901.
15. Lossos,I.S., Alizadeh,A.A., Diehn,M., Warnke,R., Thorstenson,Y., Oefner,P.J., Brown,P.O., Botstein,D. and Levy,R. (2002) Transformation of follicular lymphoma to diffuse large-cell lymphoma: alternative patterns with increased or decreased expression of c-myc and its regulated genes. *Proc. Natl Acad. Sci. USA*, **99**, 8886–8891.
16. Higgins,J.P., Shinghal,R., Gill,H., Reese,J.H., Terris,M., Cohen,R.J., Fero,M., Pollack,J.R., van de Rijn,M. and Brooks,J.D. (2003) Gene expression patterns in renal cell carcinoma assessed by complementary DNA microarray. *Am. J. Pathol*., **162**, 925–932.
17. Nielsen,T.O., West,R.B., Linn,S.C., Alter,O., Knowling,M.A., O'Connell,J.X., Zhu,S., Fero,M., Sherlock,G., Pollack,J.R. *et al*. (2002) Molecular characterisation of soft tissue tumours: a gene expression study. *Lancet*, **359**, 1301–1307.
18. Garber,M.E., Troyanskaya,O.G., Schluens,K., Petersen,S., Thaesler,Z., Pacyna-Gengelbach,M., van de Rijn,M., Rosen,G.D., Perou,C.M. and Whyte,R.I. *et al*. (2001) Diversity of gene expression in adenocarcinoma of the lung. *Proc. Natl Acad. Sci. USA*, **98**, 13784–13789.
19. van 't Veer,L.J., Dai,H., van de Vijver,M.J., He,Y.D., Hart,A.A., Mao,M., Peterse,H.L., van der Kooy,K., Marton,M.J., Witteveen,A.T. *et al*. (2002) Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, **415**, 530–536.
20. Velculescu,V.E., Zhang,L., Vogelstein,B. and Kinzler,K.W. (1995) Serial analysis of gene expression. Science, **270**, 484–487.
21. Khatri,P., Draghici,S., Ostermeier,C. and Krawetz,S. (2002) Profiling gene expression utilizing onto-express. *Genomics*, **79**, 266–270.
22. Draghici,S., Khatri,P., Martins,R.P., Ostermeier,G.C. and Krawetz,S.A. (2003) Global functional profiling of gene expression. *Genomics*, **81**, 98–104.
23. Draghici,S., Khatri,P., Bhavsar,P., Shah,A., Krawetz,S.A. and Tainsky,M.A. (2003) Onto-Tools, the toolkit of the modern biologist: Onto-Express, Onto-Compare, Onto-Design and Onto-Translate. *Nucleic Acids Res*., **31**, 3775–3781.
24. Dahlquist,K.D., Salomonis,N., Vranizan,K., Lawlor,S.C. and Conklin BR. (2002) GenMAPP, a new tool for viewing and analyzing microarray data on biological pathways. *Nature Genet*., **31**, 19–20.
25. Vilo,J., Kapushesky,M., Kemmeren,P., Sarkans,U. and Brazma,A. (2003) Expression Profiler. In Parmigiani,G., Garrett,E.S., Irizarry,R. and Zeger,S.L. (eds), *The Analysis of Gene Expression Data: Methods and Software*, Springer Verlag, New York, NY.
26. Zeeberg,B.R., Feng,W., Wang,G., Wang,M.D., Fojo,A.T., Sunshine,M., Narasimhan,S., Kane,D.W., Reinhold,W.C., Lababidi,S. *et al*. (2003) GoMiner: a resource for biological interpretation of genomic and proteomic data. *Genome Biol*., **4**, R28.
27. Zhong,S., Li,C. and Wong,W.H. (2003) ChipInfo: Software for extracting gene annotation and gene ontology information for microarray analysis. *Nucleic Acids Res*., **31**, 3483–3486.
28. Liu,G., Loraine,A.E., Shigeta,R., Cline,M., Cheng,J., Valmeekam,V., Sun,S., Kulp,D. and Siani-Rose,M.A. (2003) NetAffx: affymetrix probesets and annotations. *Nucleic Acids Res*., **31**, 82–86.
29. Tamames,J., Clark,D., Herrero,J., Dopazo,J., Blaschke,C., Fernandez,J.M., Oliveros,J.C. and Valencia,A. (2002) Bioinformatics methods for the analysis of expression arrays: data clustering and information extraction. *J. Biotechnol*., **98**, 269–283.
30. Doniger,S.W., Salomonis,N., Dahlquist,K.D., Vranizan,K., Lawlor,S.C. and Conklin,B.R. (2003) MAPPFinder: using Gene Ontology and GenMAPP to create a global gene-expression profile from microarray data. *Genome Biol*., **4**, R7.
31. Whitfield,M.L., Sherlock,G., Saldanha,A.J., Murray,J.I., Ball,C.A., Alexander,K.E., Matese,J.C., Perou,C.M., Hurt,M.M., Brown,P.O. *et al*. (2002) Identification of genes periodically expressed in the human cell cycle and their expression in tumors. *Mol. Biol. Cell*, **13**, 1977–2000.